

Licensed and Unlicensed Spectrum Allocation in Heterogeneous Networks

Zhiyi Zhou, Dongning Guo, and Michael L. Honig

Department of Electrical Engineering and Computer Science

Northwestern University, Evanston, IL 60208, USA

Abstract

In future networks, an operator may employ a wide range of access points using diverse radio access technologies (RATs) over multiple licensed and unlicensed frequency bands. This paper studies centralized user association and spectrum allocation across many access points in such a heterogeneous network (HetNet). Such centralized control is on a relatively slow timescale to allow information exchange and joint optimization over multiple cells. This is in contrast and complementary to fast timescale distributed scheduling. A queueing model is introduced to capture the lower spectral efficiency, reliability, and additional delays of data transmission over the unlicensed bands due to contention and/or listen-before-talk requirements. Two optimization-based spectrum allocation schemes are proposed along with efficient algorithms for computing the allocations. The proposed solutions are fully aware of traffic loads, network topology, as well as external interference levels in the unlicensed bands. Packet-level simulation results show that the proposed schemes significantly outperform orthogonal and full-frequency-reuse allocations under all traffic conditions.

Index Terms

Long term evolution (LTE), LTE in unlicensed spectrum (LTE-U), multiple radio access technologies, spectrum allocation, user association.

This work was supported in part by a gift from Futurewei Technologies and by the National Science Foundation under Grant No. CCF-1423040.

I. INTRODUCTION

The wireless industry has set an ambitious goal to increase the area capacity (in bits per second per square meter) by three orders of magnitude in the next five to ten years. In addition to densely deploying small cells and improving the spectral efficiency [2], another avenue is to exploit all available spectrum, which is a relatively scarce resource [3]. Future generation cellular networks are likely to involve multiple radio access technologies (RATs) over multiple frequency bands (including millimeter wave). Such RATs include Long Term Evolution (LTE), WiFi, and LTE in unlicensed spectrum (LTE-U). The prime bands today are licensed frequency bands under 3 GHz and over 500 MHz of unlicensed spectrum in the 2.4 GHz and 5 GHz frequency bands.

Current 4G cellular networks generally use regular frequency reuse patterns, including full frequency reuse and fractional frequency reuse. In the former scheme, every cell uses all available frequency bands, whereas the latter scheme is similar, except that cells use orthogonal frequency bands at the cell edge to reduce mutual interference. However, such simple methods are not as effective in emerging heterogeneous networks (HetNets), whose topologies are often highly irregular, and whose small cells have widely varying traffic conditions.

There have been many studies on spectrum allocation in cellular networks (see, e.g., [4]–[16]). Most authors formulate the allocation problem as that of deciding, for each slice of the spectrum, which access points (APs) and/or links should use it. In addition, user association was considered in [12], [13]. In general, NP-hard nonconvex integer programming problems are formulated, which typically have many local optima. Furthermore, the figures of merit used in most work are physical layer performance measures such as sum rate and outage probability. These traffic-independent metrics may not reflect a user’s relevant quality of service (QoS) in HetNets with large traffic variations in overlapping cells with complicated interference conditions.

To address the preceding issues, an alternative optimization-based framework was developed in [17] and [18] for allocating downlink spectrum resources in a HetNet. This framework allows arbitrary user association and flexible spectrum allocation, driven entirely by users’ traffic loads. This framework is well matched to emerging centralized remote processing architectures such as cloud radio access network (C-RAN) and cell-free massive multiple-input multiple-output systems [19]. In contrast to most existing work [6]–[10], which considers resource allocation on

the timescale of a frame, here the timescale of resource adaptation is conceived to be once every few seconds or minutes. This timescale is, on the one hand, fast enough for tracking aggregate traffic variations and large-scale fading, and, on the other hand, slow enough to allow information exchange and joint optimization of many APs with a large number of user equipments (UEs). Another advantage is that the spectrum resources can be assumed to be homogeneous on this timescale, namely, every segment in the same band has about the same spectral efficiency when averaged over many frames. The approach in [17] has been generalized to incorporate energy-efficient allocation via cell activation [18] as well as the effect of opportunistic scheduling on a fast timescale [20], [21].

In this work we generalize the framework of [17] and [18] to the scenario in which there are multiple RATs over multiple frequency bands. Resource allocation over multiple RATs has been studied in [22]–[24]. In [22], the authors presented a scheme for balancing licensed and unlicensed traffic in the case of a single femto user and single WiFi user. In [23], [24], joint licensed and unlicensed resource allocation algorithms were proposed for licensed-assisted access systems for throughput and energy efficiency maximization, respectively. However, the distinct characteristics of the unlicensed bands were not modeled in those papers. In our work, different queueing models are proposed to distinguish the characteristics of licensed and unlicensed bands. The multiple-band allocation problem is formulated as a bi-convex optimization problem. A conservative allocation scheme is first proposed following the approach of [17] and [18]. This is followed by a utilization-dependent allocation scheme which incorporates the utilizations of the APs into the formulation to more accurately account for dynamic inter-cell interference. An iterative algorithm is designed to solve the allocation problem with manageable computational complexity for small systems. In each step, we solve a convex problem with a unique optimal solution.

Another feature of this work is that the packet length is allowed to have a general probability distribution. In prior work [17], [18], [20] and [21], the packet length is assumed to be exponentially distributed to yield a simple analytic form in the objective representing the QoS. We show that the proposed formulation and algorithm apply to general packet length distributions and hence a broader class of traffic conditions.

To evaluate the performance, packet-level simulations are carried out. It is shown that both the conservative allocation scheme and the utilization-based allocation scheme significantly reduce

the average packet delay in the heavy traffic regime compared to orthogonal allocation and full-frequency-reuse allocation. The large performance gain is observed mainly because the proposed schemes are traffic-aware and also exploit the particular characteristics of each RAT. In addition, the utilization-dependent allocation scheme attains the best performance over all traffic conditions due to its accurate modeling of the dynamic interference among APs.

One practical application of this work is the emerging C-RAN which allows many remote radio function units connect to a centralized network controller. The total overhead for the network controller to perform the proposed resource allocation scheme includes collecting the spectral efficiencies of all links and users' traffic information. Since the timescale of resource adaptation is considered to be once every a few seconds or minutes, the overhead is quite small. For example, the rate for sending 30,000 parameters (16 bits each) every minute is only 8 kilobits per second (kbps).

The remainder of this paper is organized as follows. The system model is introduced in Section II. An optimization problem using the conservative allocation is formulated in Section III. A utilization-dependent allocation scheme is presented in Section IV. The extension to general packet length distributions is given in Section V. Simulation results are presented in Section VI and concluding remarks are given in Section VII. All technical proofs are relegated to the appendices.

II. SYSTEM MODEL

In this section, we introduce models for user traffic, spectrum allocation, and link throughput. The models extend those in [18] to accommodate multiple RATs. A new queueing model is then introduced for traffic over unlicensed bands.

A. Spectrum Allocation

We consider the downlink of a heterogeneous network consisting of n APs and many UEs. Without loss of generality, suppose each AP can use all m different RATs, with each RAT on a separate frequency band. Since UEs located near each other often have similar channel conditions, these UEs can be treated as a group on the slow timescale. This is a generalization of the extreme case where each UE group contains only one single UE. Moreover, it suffices to carry out the slow timescale allocation to the UE groups rather than individual UEs.

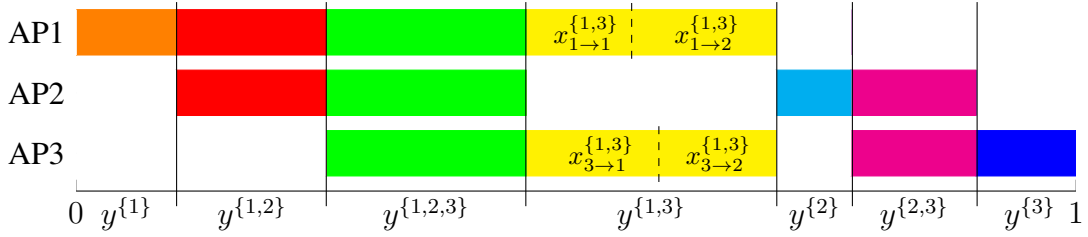


Fig. 1. Illustration of all patterns of a 3-AP network. The allocations to 2 users are revealed under pattern $\{1, 3\}$.

Denote the set of all APs by $\mathcal{N} = \{1, \dots, n\}$, the set of all user groups by $\mathcal{K} = \{1, \dots, k\}$, and the set of RATs by $\mathcal{M} = \{1, \dots, m\}$. RAT l employs its separate homogeneous spectrum of bandwidth $w^{(l)}$. We allow arbitrary association so that each AP can simultaneously serve any subset of UE groups and each UE can be simultaneously served by any subset of APs. Furthermore, we allow flexible resource allocation in that each AP-UE link can use an arbitrary number of RATs, where each RAT uses an arbitrary (possibly discontinuous) subset of the available spectrum. Despite the enormous number of possibilities, we will show that the actual AP-UE association and spectrum allocation is extremely sparse.

The key to total spectrum agility is the notion of *pattern* [17], [18]. In general, a pattern simply refers to a subset of transmitters. A time-frequency resource is said to be reserved for pattern \mathcal{A} if the resource is to be shared by transmitters in \mathcal{A} .¹ In the downlink, a pattern \mathcal{A} is a subset of \mathcal{N} , and all APs in \mathcal{A} have access to the time-frequency resources associated with pattern \mathcal{A} . Assuming known transmit power spectral densities (PSDs), the pattern of a resource determines the signal-to-interference-plus-noise ratio (SINR) and hence the spectral efficiencies of all links over the resource. The allocation problem can then be formulated as how to divide the spectrum of each RAT among all $2^n - 1$ nonempty patterns. To illustrate the concept of pattern with an example as shown in Fig. 1, suppose there are three APs operating over one frequency band. The spectrum can be divided into $2^3 - 1 = 7$ segments, where one segment is used by AP 1 exclusively (the pattern is $\{1\}$), a second is used by AP 2 exclusively (the pattern is $\{2\}$), a third is used by AP 3 exclusively (the pattern is $\{3\}$) and the remaining four segments

¹The notion of pattern finds its root in the concept of independent set defined in the special case where the network is described by a conflict graph, and where nodes/links in the independent set share the same resources [25].

include three shared by the two APs (the patterns are $\{1, 2\}$, $\{2, 3\}$, and $\{1, 3\}$, respectively) as well as one segment shared by all three APs (the pattern is $\{1, 2, 3\}$). If the transmit pattern of a certain spectrum resource is $\{1, 2\}$, i.e., both APs transmit, then the spectral efficiencies of all links are determined as discussed below.

A network controller collects traffic load and channel/interference information from all the APs. Because the time period of slow timescale allocation is much longer than the channel coherence time, the channel conditions are modeled using path loss and the statistics of fading. Given the average channel and traffic conditions, the task of the central controller is to determine which spectrum segments to allocate to each AP, and subsequently, which sub-segments to allocate to specific UEs associated with that AP. To be precise, we need to solve the following three subproblems:

- 1) Decide which RAT or RATs should be used to serve each UE group.
- 2) Allocate the bandwidth across all 2^n patterns for each RAT. This is denoted by a $2^n \times m$ -tuple: $y = (y^{\mathcal{A},l})_{\mathcal{A} \subset \mathcal{N}, l \in \mathcal{M}}$, where $y^{\mathcal{A},l} \in [0, 1]$ is the fraction of spectrum assigned to RAT l shared by APs in set \mathcal{A} . Clearly,

$$\sum_{\mathcal{A} \subset \mathcal{N}} y^{\mathcal{A},l} = 1, \quad \forall l \in \mathcal{M} \quad (1)$$

and any efficient allocation would not use the empty pattern, so that $y^{\emptyset,l} = 0$, for every $l \in \mathcal{M}$.

- 3) For every pattern $\mathcal{A} \subset \mathcal{N}$, for every AP $i \in \mathcal{A}$, divide the bandwidth $y^{\mathcal{A},l}$ to the UE groups. Denote the bandwidth allocated to the link $i \rightarrow j$ over pattern \mathcal{A} under RAT l as $x_{i \rightarrow j}^{\mathcal{A},l}$. As shown in Fig. 1, $y^{\{1,3\}}$ (colored yellow) is further divided by AP 1 into two parts to serve UE group 1 and UE group 2, respectively, while AP 3 divides the same shared spectrum differently. Then we have:

$$\sum_{j \in \mathcal{K}} x_{i \rightarrow j}^{\mathcal{A},l} \leq y^{\mathcal{A},l}, \quad \forall i \in \mathcal{A}, \mathcal{A} \subset \mathcal{N}, l \in \mathcal{M}. \quad (2)$$

The user association is indirectly determined by the amount of spectrum resources assigned by each AP to each group. Specifically, UE j is assigned to AP i over RAT l if and only if $x_{i \rightarrow j}^{\mathcal{A},l} > 0$ for some pattern \mathcal{A} .

B. From Spectrum to Transmission Rates

An important measure of user QoS is the average packet delay in the system. The delay is determined by the packet arrival statistics and the service rates. The instantaneous service rate of a UE group's queue in turn depends on the spectral efficiencies along with the bandwidths allocated to that specific UE group. For simplicity, it is assumed that when AP i transmits over RAT l , it employs all patterns available to it and applies a flat PSD $p_i^{(l)}$ over the allocated spectrum. At any frequency designated for RAT l , the instantaneous spectral efficiency achievable by the link from AP i to UE j depends on the set of active APs $\mathcal{A} \subset \mathcal{N}$ using that frequency. Let this (spectral) efficiency be denoted by $s_{i \rightarrow j}^{A,l}$. Clearly, $s_{i \rightarrow j}^{A,l} = 0$ if $i \notin \mathcal{A}$. Moreover, the spectral efficiency decreases as more APs become active, i.e., $s_{i \rightarrow j}^{A,l} \geq s_{i \rightarrow j}^{B,l}$ if $\mathcal{A} \subset \mathcal{B}$. On the slow timescale, the spectral efficiencies are either known *a priori* or can be measured and sent to the central controller. For later convenience, we normalize the spectral efficiency over RAT l using factor $L/w^{(l)}$ (bits/second/Hz) so that the unit of $s_{i \rightarrow j}^{A,l}$ is packets/second.

For concreteness in obtaining numerical results, we use Shannon's formula to arrive at the expression for link efficiencies:

$$s_{i \rightarrow j}^{A,l} = \frac{\alpha^{(l)} w^{(l)} \mathbb{1}(i \in \mathcal{A})}{L} \log_2 \left(1 + \frac{p_i^{(l)}}{I_{i \rightarrow j}^{A,l}} \right) \text{ packets/s,} \quad (3)$$

where L is the average packet length in bits,² $\alpha^{(l)} \in (0, 1]$ is the discount factor for RAT l , $\mathbb{1}(i \in \mathcal{A}) = 1$ if $i \in \mathcal{A}$ and $\mathbb{1}(i \in \mathcal{A}) = 0$ otherwise, and $I_{i \rightarrow j}^{A,l}$ is the total noise plus interference PSD from other APs in \mathcal{A} to UE j , which depends on their transmit PSDs and path loss. The discount factor of a licensed band is typically closer to 1 than that of an unlicensed band due to external interference from other operators in the unlicensed band. The discount factor is not crucial and is merely included for flexibility of the model. The average effect of small-scale fading can be included by considering the ergodic capacity in lieu of (3), which does not change the main developments of this paper.

Over a resource reserved for pattern \mathcal{A} , only the subset of APs in \mathcal{A} with data to transmit will be using the resource at any given time. In general, the instantaneous service rate depends on

²Packet lengths are assumed to be i.i.d. with exponential distribution; later we will show that the proposed methods can be applied to general packet length distributions.

the set of transmitting APs and the rate adaptation scheme. Two service rate models are adopted as described in the following.

Under the conservative allocation introduced in [17], if some of RAT l 's spectrum resources reserved for pattern \mathcal{A} are allocated to link $i \rightarrow j$, then AP i transmits UE j 's packets at spectral efficiency $s_{i \rightarrow j}^{\mathcal{A},l}$ using RAT l over pattern \mathcal{A} . This rate is achievable even if all APs in \mathcal{A} have data to transmit (hence the name conservative). This is equivalent to assuming that other APs' traffic is always backlogged. The rate contributed by spectrum reserved for pattern \mathcal{A} under RAT l is the product of the spectral efficiency and the bandwidth: $s_{i \rightarrow j}^{\mathcal{A},l} x_{i \rightarrow j}^{\mathcal{A},l}$. The total service rate for the queue of UE j over RAT l is the sum rate of all APs and patterns, expressed as:

$$r_j^{(l)} = \sum_{\mathcal{A} \subset \mathcal{N}} \sum_{i \in \mathcal{N}} s_{i \rightarrow j}^{\mathcal{A},l} x_{i \rightarrow j}^{\mathcal{A},l} \text{ packets/s.} \quad (4)$$

The advantage of the conservative allocation is that there is no need for the scheduler (or whichever unit that performs rate adaptation) to know the state of the other APs included in the pattern. However, the conservative rate (4) is a lower bound on the actual rate because the interference is overestimated.

To adapt the allocation to the actual interference pattern, we introduce a new model in this work, referred to as the *utilization-dependent model*. Under this model, the service rate contributed by the spectrum reserved for pattern \mathcal{A} depends on the subset of transmitting APs in \mathcal{A} . Specifically, the service rate for UE group j over RAT l when the set of active interfering APs is \mathcal{I} is expressed as:

$$r_j^{\mathcal{I},l} = \sum_{\mathcal{A} \subset \mathcal{N}} \sum_{i \in \mathcal{N}} s_{i \rightarrow j}^{\mathcal{A} \cap \mathcal{I},l} x_{i \rightarrow j}^{\mathcal{A},l} \text{ packets/s.} \quad (5)$$

C. Queueing Model

Without loss of generality, we restrict the subsequent investigation to the scenario with two RATs, where one RAT is over the licensed band (conceived as LTE) and the other is over the unlicensed band (conceived as LTE-U). This can be easily generalized to more than two RATs.

The traffic to UE group j is modeled as an independent Poisson point process with rate λ_j packets per second. The packet lengths are independent random variables whose average is L bits. Packets intended for each UE group are transmitted according to the first-in-first-out (FIFO) discipline. The buffer in each queue is assumed to be unlimited for simplicity. The traffic load of



Fig. 2. Illustration of queueing model of LTE-U.

UE group j is further divided into two streams served by the two RATs, respectively. This may be implemented by dividing the group of UEs for association with different RATs, so that each UE is only served by one RAT. The resulting packet arrival rates of two streams are denoted as $\lambda_j^{(1)}$ and $\lambda_j^{(2)}$, respectively, which are variables to be optimized subject to the constraint:

$$\lambda_j^{(1)} + \lambda_j^{(2)} \geq \lambda_j. \quad (6)$$

For each RAT, the corresponding k UE groups form a system of k interactive queues where the instantaneous service rate of each queue may depend on which other queues are empty.

The physical and/or medium access control layers of LTE-U are designed to facilitate coexistence with other RATs in unlicensed bands, such as WiFi.³ In [27], a listen-before-talk scheme was proposed in which carrier sensing is embedded in a deterministic portion of a LTE subframe. In [28], the authors proposed the use of LTE uplink power control to improve LTE/WiFi coexistence. References [29] and [30] propose to enable LTE/WiFi coexistence by muting LTE transmission on certain subframes following a pre-determined pattern.

In this paper, we assume that LTE-U has a listen-before-talk feature, which is likely to be a dominant mode in emerging LTE-U standards [26]. In this mode, an AP operating LTE-U with data to send first performs carrier sensing before its transmission. We use queues with vacation and nonexhaustive service [31] to model such a mechanism. The queueing scheme is depicted in Fig. 2. Specifically, the server of queue j will take a single vacation after completing the service of each packet. The vacation duration V_j (in seconds) is a random variable following a certain distribution. The mean and variance of the vacation time depend on the level of local

³Challenges of such coexistence are discussed in [26].

interference,⁴ such as interference from WiFi users. The higher the interference level, the more time a queue has to wait before being served.

III. QUEUEING DELAYS AND A CONSERVATIVE ALLOCATION SCHEME

In this section, we adopt the conservative service rate model (4) to develop a scheme for allocating licensed and unlicensed spectrum. For each RAT, the corresponding k UE groups form k independent $M/M/1$ queues, because the service rate of each queue is independent of the states of other queues under the model (4). Let RAT 1 represent LTE and RAT 2 represent LTE-U.

Under LTE, the average packet delay (in seconds/packet) of UE group j is given by the average delay for the $M/M/1$ queue:

$$t_j^{(1)} = \frac{1}{\left(r_j^{(1)} - \lambda_j^{(1)}\right)^+}, \quad (7)$$

where $(x)^+ = x$ if $x > 0$ and $(x)^+ = 0$ if $x \leq 0$. If $r_j^{(1)} \leq \lambda_j^{(1)}$, the queueing delay is infinite, i.e., the queue becomes unstable. It is important to note that the right hand side of (7) is convex in the pair $(r_j^{(1)}, \lambda_j^{(1)})$ on \mathbb{R}^2 .

Under LTE-U, the k UE groups form k independent $M/M/1$ queues with single vacation. We have:

Proposition 1: The average packet delay (in seconds/packet) of a queue with Poisson arrival of rate λ and a single server of rate r with single vacation V whose expected squared duration is $\nu = E[V^2]$ is given by:

$$t = \frac{2 + r\lambda\nu}{2(r - \lambda)^+}. \quad (8)$$

For fixed ν , the function defined by (8) is convex in the pair (r, λ) on \mathbb{R}^2 .

The proof of Proposition 1 is given in Appendix A.

⁴Although coexistence issues in the unlicensed band are very important over fast timescales, over slow timescales, effective interference levels and associated packet delays suffice to characterize how LTE-U interacts with other devices in the unlicensed band.

The joint user association and conservative spectrum allocation problem is formulated as P1:

$$\underset{\mathbf{r}, \mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \mathbf{t}}{\text{minimize}} \quad \frac{1}{\sum_{j \in \mathcal{K}} \lambda_j} \sum_{j \in \mathcal{K}} (\lambda_j^{(1)} t_j^{(1)} + \lambda_j^{(2)} t_j^{(2)}) \quad (\text{P1a})$$

$$\text{subject to} \quad t_j^{(1)} = \frac{1}{\left(r_j^{(1)} - \lambda_j^{(1)}\right)^+}, \quad j \in \mathcal{K} \quad (\text{P1b})$$

$$t_j^{(2)} = \frac{2 + r_j^{(2)} \lambda_j^{(2)} \nu_j}{2 \left(r_j^{(2)} - \lambda_j^{(2)}\right)^+}, \quad j \in \mathcal{K} \quad (\text{P1c})$$

$$r_j^{(l)} = \sum_{\mathcal{A} \subset \mathcal{N}} \sum_{i \in \mathcal{N}} s_{i \rightarrow j}^{\mathcal{A}, l} x_{i \rightarrow j}^{\mathcal{A}, l}, \quad j \in \mathcal{K}, l \in \{1, 2\} \quad (\text{P1d})$$

$$\lambda_j^{(1)} + \lambda_j^{(2)} \geq \lambda_j, \quad j \in \mathcal{K} \quad (\text{P1e})$$

$$y^{\mathcal{A}, l} \geq \sum_{j \in \mathcal{K}} x_{i \rightarrow j}^{\mathcal{A}, l}, \quad l \in \{1, 2\}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N} \quad (\text{P1f})$$

$$\sum_{\mathcal{A} \subset \mathcal{N}} y^{\mathcal{A}, l} = 1, \quad l \in \{1, 2\} \quad (\text{P1g})$$

$$x_{i \rightarrow j}^{\mathcal{A}, l} \geq 0, \quad j \in \mathcal{K}, l \in \{1, 2\}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N}. \quad (\text{P1h})$$

The objective (P1a) is the average packet delay of all queues over the entire network. (P1b) is the average packet delay of each UE group served by LTE as derived in (7). (P1c) is the average packet delay of each UE group served by LTE-U as derived in (8). (P1d) is the service rate of each divided UE group using the conservative model as given in (4). (P1e) is the total traffic constraint for each user type (6). (P1f) guarantees the consistency of the spectrum allocation as given in (2). (P1g) constrains the total bandwidth of each RAT to be one unit. (P1h) precludes negative bandwidth. The variables in P1 are $\mathbf{r}, \mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \mathbf{t}$ which are the vector forms of $(r_j^{(l)})_{j \in \mathcal{K}, l \in \{1, 2\}}$, $(x_{i \rightarrow j}^{\mathcal{A}, l})_{j \in \mathcal{K}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N}, l \in \{1, 2\}}$, $(y^{\mathcal{A}, l})_{\mathcal{A} \subset \mathcal{N}, l \in \{1, 2\}}$, $(\lambda_j^{(l)})_{j \in \mathcal{K}, l \in \{1, 2\}}$, $(t_j^{(l)})_{j \in \mathcal{K}, l \in \{1, 2\}}$, respectively.

P1 is a bi-convex optimization problem because when variable $\boldsymbol{\lambda}$ (resp. \mathbf{x}) is fixed, all constraints are linear and the objective is a linear combination of convex functions in \mathbf{x} (resp. $\boldsymbol{\lambda}$). Therefore, P1 can be solved by alternating optimization over $\boldsymbol{\lambda}$ and \mathbf{x} . Since the sequence of objective values obtained at each iteration is lower-bounded and non-increasing, the objective converges although it may not be a global minimum. As shown in Section VI, the alternating method achieves good performance.

Theorem 1: The (global) minimum average delay can be achieved by a sparse allocation,

where the spectrum of each RAT is divided into at most k segments. That is, there exists an optimal solution that satisfies

$$|\{\mathcal{A} \mid y^{\mathcal{A},l} > 0, \mathcal{A} \subset \mathcal{N}\}| \leq k, \quad \forall l \in \mathcal{M} \quad (9)$$

where $|\cdot|$ denotes the cardinality of a set. The same applies to the delay achieved by the alternating optimization method.

Theorem 1 is proved in Appendix B. The theorem guarantees that although the number of all possible patterns grows exponentially with the number of APs in the network, using a small number of patterns achieves the optimal performance.

Theorem 2: Assume the channel gains are jointly continuous random variables. The (global) minimum average delay can be achieved with at most $n - 1$ user groups served by multiple APs on each RAT. That is, there exists an optimal solution that satisfies

$$\left| \{j \mid x_{i_1 \rightarrow j}^{\mathcal{A}_1, l}, x_{i_2 \rightarrow j}^{\mathcal{A}_2, l} > 0, \text{ for some } i_1, i_2 \in \mathcal{N} \text{ and } \mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{N}\} \right| \leq n - 1, \quad \forall l \in \mathcal{M}. \quad (10)$$

The same applies to the delay achieved by the alternating optimization method.

Theorem 2 is proved in Appendix C. The theorem guarantees that although we allow a UE group to be served by multiple APs, most UE groups will be associated with only one AP in the optimal solution.

IV. A UTILIZATION-DEPENDENT ALLOCATION SCHEME

In this section, we adopt the utilization-dependent service rate model (5). Unlike in the conservative scheme, we now let each AP adapt its transmission rates to the instantaneous set of active interfering APs. Thus the service rate is in general higher than the conservative rate.

In a stable interactive queueing system, each AP transmits (over each RAT) for a fraction of the time, referred to as the *utilization*. Denote the utilization of AP i over RAT l as $\rho_i^{(l)} \in [0, 1]$. The analysis of interactive queueing system is difficult. As an approximation, we assume that different APs transmit independently over each RAT. The probability of an active interfering AP set \mathcal{I} over RAT l is then:

$$p^{\mathcal{I}, l} = \left(\prod_{i \in \mathcal{I}} \rho_i^{(l)} \right) \left(\prod_{i' \notin \mathcal{I}} (1 - \rho_{i'}^{(l)}) \right). \quad (11)$$

When UE group j is served, its instantaneous service rate is one of 2^n possible values depending on the set of active APs. Thus we use a certain “average” of 2^n independent $M/M/1$ queues to

approximate the queueing behavior of UE group j with interactive queues. Such an approximation is reasonable under the premise that for any AP, the influence of other APs is adequately represented by their steady state probability distribution. Under each RAT, the average delay of UE group j is calculated as the expected delay over 2^n possible rates:

$$t_j^{(1)} = \sum_{\mathcal{I} \subset \mathcal{N}} \frac{p^{\mathcal{I},1}}{(r_j^{\mathcal{I},1} - \lambda_j^{(1)})_+} \quad (12a)$$

$$t_j^{(2)} = \sum_{\mathcal{I} \subset \mathcal{N}} \frac{p^{\mathcal{I},2}(2 + r_j^{\mathcal{I},2}\lambda_j^{(2)}\nu_j)}{2(r_j^{\mathcal{I},2} - \lambda_j^{(2)})_+}. \quad (12b)$$

The utilization of UE group j is calculated as its expected utilization over all possible sets of interfering APs. Specifically, when the active set of APs is \mathcal{I} , the fraction of time that group j is served is $\frac{\lambda_j^{(l)}}{r_j^{\mathcal{I},l}}$, the average utility is obtained as:

$$\sigma_j^{(l)} = \sum_{\mathcal{I} \subset \mathcal{N}} p^{\mathcal{I},l} \frac{\lambda_j^{(l)}}{r_j^{\mathcal{I},l}}, \quad l \in \{1, 2\}. \quad (13)$$

Since AP i may serve multiple UE groups over each RAT, the utilization of AP i depends on the utilization of its associated UE groups. We approximate the utilization of AP i over RAT l as its average utilization over the spectrum used. Specifically, the average amount of spectrum used by AP i to serve its UEs is $\sum_{j \in \mathcal{K}} \sigma_j^{(l)} \sum_{\mathcal{A}: i \in \mathcal{A}} x_{i \rightarrow j}^{\mathcal{A},l}$. Hence AP i 's utilization is approximated as

$$\rho_i^{(l)} = \frac{1}{\sum_{\mathcal{A}: i \in \mathcal{A}} y^{\mathcal{A},l}} \sum_{\mathcal{A}: i \in \mathcal{A}} \sum_{j \in \mathcal{K}} \sigma_j^{(l)} x_{i \rightarrow j}^{\mathcal{A},l}, \quad (14)$$

where $\sum_{\mathcal{A}: i \in \mathcal{A}} y^{\mathcal{A},l}$ is the total bandwidth used by AP i over RAT l .

With the above approximation, the joint user association and utilization-dependent spectrum

allocation problem is formulated as P2:

$$\begin{aligned} & \underset{\mathbf{r}, \mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \mathbf{t}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \mathbf{p}}{\text{minimize}} && \frac{1}{\sum_{j \in \mathcal{K}} \lambda_j} \sum_{j \in \mathcal{K}} (\lambda_j^{(1)} t_j^{(1)} + \lambda_j^{(2)} t_j^{(2)}) \end{aligned} \quad (\text{P2a})$$

$$\text{subject to} \quad t_j^{(1)} = \sum_{\mathcal{I} \subset \mathcal{N}} \frac{p^{\mathcal{I},1}}{\left(r_j^{\mathcal{I},1} - \lambda_j^{(1)}\right)^+}, \quad j \in \mathcal{K} \quad (\text{P2b})$$

$$t_j^{(2)} = \sum_{\mathcal{I} \subset \mathcal{N}} \frac{p^{\mathcal{I},2} (2 + r_j^{\mathcal{I},2} \lambda_j^{(2)} \nu_j)}{2 \left(r_j^{\mathcal{I},2} - \lambda_j^{(2)}\right)^+}, \quad j \in \mathcal{K} \quad (\text{P2c})$$

$$r_j^{\mathcal{I},l} = \sum_{\mathcal{A} \subset \mathcal{N}} \sum_{i \in \mathcal{N}} s_{i \rightarrow j}^{\mathcal{A} \cap \mathcal{I},l} x_{i \rightarrow j}^{\mathcal{A},l}, \quad \mathcal{I} \subset \mathcal{N}, j \in \mathcal{K}, l \in \{1, 2\} \quad (\text{P2d})$$

$$\lambda_j^{(1)} + \lambda_j^{(2)} \geq \lambda_j, \quad j \in \mathcal{K} \quad (\text{P2e})$$

$$y^{\mathcal{A},l} \geq \sum_{j \in \mathcal{K}} x_{i \rightarrow j}^{\mathcal{A},l}, \quad l \in \{1, 2\}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N} \quad (\text{P2f})$$

$$\sum_{\mathcal{A} \subset \mathcal{N}} y^{\mathcal{A},l} = 1, \quad l \in \{1, 2\} \quad (\text{P2g})$$

$$x_{i \rightarrow j}^{\mathcal{A},l} \geq 0, \quad j \in \mathcal{K}, l \in \{1, 2\}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N} \quad (\text{P2h})$$

$$p^{\mathcal{I},l} = \left(\prod_{i \in \mathcal{I}} \rho_i^{(l)} \right) \left(\prod_{i' \notin \mathcal{I}} (1 - \rho_{i'}^{(l)}) \right), \quad l \in \{1, 2\}, \mathcal{I} \subset \mathcal{N} \quad (\text{P2i})$$

$$\rho_i^{(l)} = \frac{1}{\sum_{\mathcal{A}: i \in \mathcal{A}} y^{\mathcal{A},l}} \sum_{\mathcal{A}: i \in \mathcal{A}} \sum_{j \in \mathcal{K}} \sigma_j^{(l)} x_{i \rightarrow j}^{\mathcal{A},l}, \quad l \in \{1, 2\}, i \in \mathcal{N} \quad (\text{P2j})$$

$$\sigma_j^{(l)} \geq \sum_{\mathcal{I} \subset \mathcal{N}} p^{\mathcal{I},l} \frac{\lambda_j^{(l)}}{r_j^{\mathcal{I},l}}, \quad l \in \{1, 2\}, j \in \mathcal{K}. \quad (\text{P2k})$$

The objective (P2a) is also average packet delay of all queues of the entire network. (P2d) is the service rate of each divided UE group using utilization model as in (5). The variables in P2 are $\mathbf{r}, \mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}, \mathbf{t}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \mathbf{p}$ which are of the vector forms of $(r_j^{\mathcal{I},l})_{\mathcal{I} \subset \mathcal{N}, j \in \mathcal{K}, l \in \{1,2\}}$, $(x_{i \rightarrow j}^{\mathcal{A},l})_{j \in \mathcal{K}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N}, l \in \{1,2\}}$, $(y^{\mathcal{A},l})_{\mathcal{A} \subset \mathcal{N}, l \in \{1,2\}}$, $(\lambda_j^{(l)})_{j \in \mathcal{K}, l \in \{1,2\}}$, $(t_j^{(l)})_{j \in \mathcal{K}, l \in \{1,2\}}$, $(\sigma_j^{(l)})_{j \in \mathcal{K}, l \in \{1,2\}}$, $(\rho_i^{(l)})_{i \in \mathcal{N}, l \in \{1,2\}}$, and $(p^{\mathcal{I},l})_{\mathcal{I} \subset \mathcal{N}, l \in \{1,2\}}$, respectively. P2 would be equivalent to P1 if we let $p^{\mathcal{N},l} = 1$, $p^{\mathcal{I},l} = 0$ for every $\mathcal{I} \neq \mathcal{N}$, and $\sigma_j^{(l)} = 1$ for every l and j . (This is a feasible suboptimal solution.) Unlike P1, P2 is not bi-convex because new variables and nonlinear constraints are introduced. However, when fixing the variables $\boldsymbol{\sigma}, \boldsymbol{\rho}$ and \mathbf{p} , it becomes bi-convex. Therefore, we divide P2 into two subproblems and solve them alternately to update all the variables.

Algorithm 1 Update σ , ρ and p

Step 1. Initialization: set initial utilization of UE groups $\sigma(0)$;

Step 2. Update: $\sigma_j^{(l)}(m) \leftarrow \sum_{\mathcal{I} \subset \mathcal{N}} p^{\mathcal{I},l}(m-1) \frac{\lambda_j^{(l)}}{r_j^{\mathcal{I},l}}, \quad l \in \{1, 2\}, j \in \mathcal{K}$

$$\rho_i^{(l)}(m) \leftarrow \frac{1}{\sum_{\mathcal{A}: i \in \mathcal{A}} y^{\mathcal{A},l}} \sum_{\mathcal{A}: i \in \mathcal{A}} \sum_{j \in \mathcal{K}} \sigma_j^{(l)}(m) x_{i \rightarrow j}^{\mathcal{A},l}, \quad l \in \{1, 2\}, i \in \mathcal{N}$$

$$p^{\mathcal{I},l}(m) \leftarrow \left(\prod_{i \in \mathcal{I}} \rho_i^{(l)}(m) \right) \left(\prod_{i' \notin \mathcal{I}} (1 - \rho_{i'}^{(l)}(m)) \right), \quad l \in \{1, 2\}, \mathcal{I} \subset \mathcal{N}$$

$$m \leftarrow m + 1;$$

Step 3. if $\|\sigma(m) - \sigma(m-1)\| \geq \epsilon$, where ϵ is a small number, go back to Step 2; otherwise stop and obtain

$$\sigma_j^{(l)} \leftarrow \sigma_j^{(l)}(m), \rho_i^{(l)} \leftarrow \rho_i^{(l)}(m), p^{\mathcal{I},l} \leftarrow p^{\mathcal{I},l}(m), \quad l \in \{1, 2\}, j \in \mathcal{K}, i \in \mathcal{N}, \mathcal{I} \subset \mathcal{N}.$$

Algorithm 2 Iterative algorithm for solving P2

Initialization: $x \leftarrow 0; x' \leftarrow 1; \sigma_j^{(l)} \leftarrow 1, \forall j \in \mathcal{K}, l \in \{1, 2\}; \rho_i^{(l)} \leftarrow 1, \forall i \in \mathcal{N}, l \in \{1, 2\}; p^{\mathcal{N},l} \leftarrow 1, p^{\mathcal{I},l} \leftarrow 0, \forall \mathcal{I} \subset \mathcal{N}, \mathcal{I} \neq \mathcal{N}.$

while $\|x - x'\| > \epsilon$ **do**

1. $x' \leftarrow x$;
2. Update r, x, y, λ, t by solving P3;
3. Update σ, ρ, p by using Algorithm 1.

end while

Given σ, ρ and p , we update r, x, y, λ, t by solving subproblem P3:

$$\underset{r, x, y, \lambda, t}{\text{minimize}} \quad \frac{1}{\sum_{j \in \mathcal{K}} \lambda_j} \sum_{j \in \mathcal{K}} (\lambda_j^{(1)} t_j^{(1)} + \lambda_j^{(2)} t_j^{(2)}) \quad (\text{P3a})$$

$$\text{subject to} \quad \sum_{\mathcal{I} \subset \mathcal{N}} p^{\mathcal{I},l} \frac{\lambda_j^{(l)}}{r_j^{\mathcal{I},l}} \leq \sigma_j^{(l)}, \quad l \in \{1, 2\}; j \in \mathcal{K} \quad (\text{P3b})$$

(P2b) – (P2h).

(P3b) constraints the utilization of UE groups. The structure of subproblem P3 is similar to that of P1. It can be solve by alternating optimization over λ and x .

Given r, x, y, λ, t , we update σ, ρ and p . There are $2(2^n + n + k)$ variables and equations. To solve them with low complexity, Algorithm 1 updates σ, ρ and p iteratively as in [20], [21].

Here m denotes the iteration. Convergence of the algorithm can be established similarly as in [20], [21]. In addition, the solution obtained by Algorithm 1 is feasible for (P2i)-(P2k).

After solving two subproblems alternately, P2 can be solved using Algorithm 2. Each step in Algorithm 2 is an easier problem. Convergence of Algorithm 2 can also be similarly established as in [20], [21]. Therefore, P2 can be solved iteratively with low complexity.

V. EXTENSION

So far, the packet lengths are assumed to be i.i.d with exponential distribution, which results in exponentially distributed service time for each packet. In this section, we show that the proposed schemes and algorithms are also applicable to general packet length distributions as long as the first and second moments of the service time can be written in the form:

$$E[X] = \frac{\beta}{r}, \quad (15a)$$

$$E[X^2] = \frac{\eta}{r^2}, \quad (15b)$$

where r is the service rate and β and η are positive numbers. A wide class of distributions have such characteristics. For example, $\beta = 1$ and $\eta = 2$ correspond to exponential service time; $\beta = 1$ and $\eta = 1$ correspond to constant service time. In general, the queues in the system are not $M/M/1$ queues. In contrast to (7) and (8), the formulas for calculating the average packet delay of UE group j become:

$$\hat{t}_j^{(1)} = \frac{(\frac{1}{2}\eta - \beta^2)\lambda_j^{(1)} + \beta r_j^{(1)}}{r_j^{(1)} \left(r_j^{(1)} - \beta \lambda_j^{(1)} \right)^+}, \quad (16a)$$

$$\hat{t}_j^{(2)} = \frac{(\frac{1}{2}\eta - \beta^2)\lambda_j^{(2)} + \beta r_j^{(2)} + \frac{1}{2}\nu_j \lambda_j^{(2)} \left(r_j^{(2)} \right)^2}{r_j^{(2)} \left(r_j^{(2)} - \beta \lambda_j^{(2)} \right)^+}. \quad (16b)$$

Then the objective function in P1 is rewritten as:

$$\hat{U} = \sum_{j \in \mathcal{K}} (\lambda_j^{(1)} \hat{t}_j^{(1)} + \lambda_j^{(2)} \hat{t}_j^{(2)}). \quad (17)$$

Proposition 2: The new objective function (17) is bi-convex in λ and r .

Proposition 2 is proved in Appendix D. Due to Proposition 2, the techniques proposed in Section III and Section IV still apply to the problem of user association and spectrum allocation with general packet length distribution .

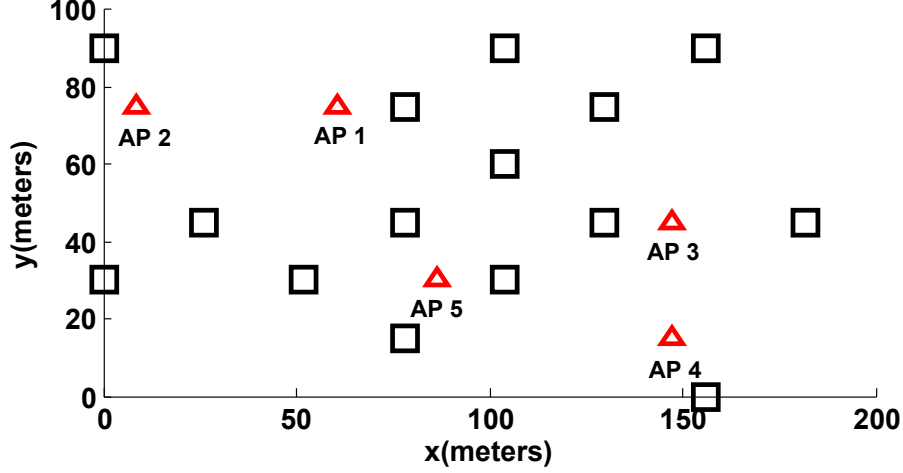


Fig. 3. The topology of the 5-AP network.

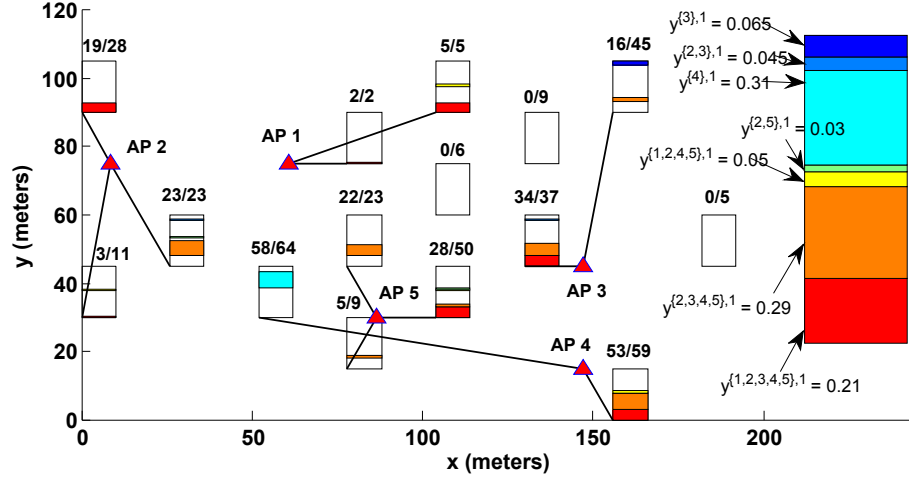
TABLE I
PARAMETER VALUES.

Parameters	Value/Function
AP transmit power	23 dBm
Total bandwidth on each RAT	10 MHz
Average packet length	0.5 Mbits
AP to UE pathloss	$140.7 + 36.7 \log_{10}(R)$

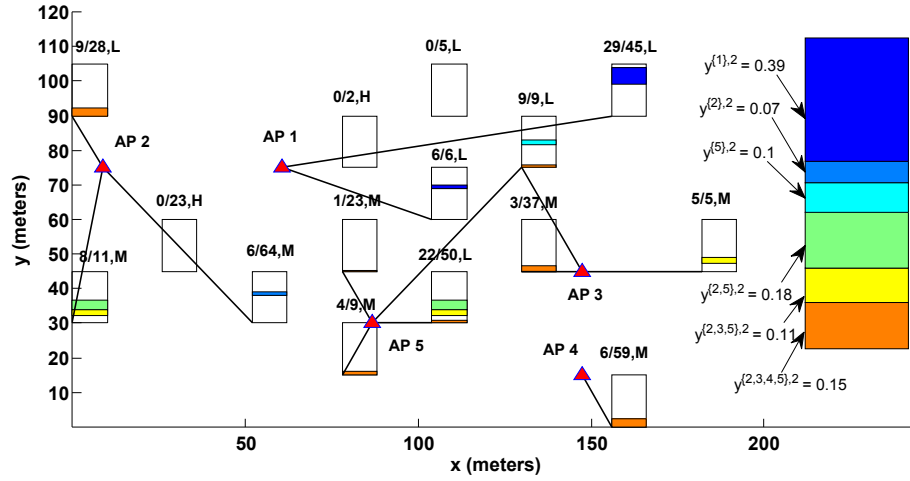
VI. NUMERICAL RESULTS

Simulations are carried out using the network topology depicted in Fig. 3. The HetNet is deployed over a $100 \times 200 \text{ m}^2$ area. Five APs, denoted by triangles, and 15 UE groups, denoted by the squares, are randomly dropped in the area. The spectral efficiency is calculated using (3) assuming $\alpha^{(1)} = 1$ and $\alpha^{(2)} = 1/2$ with a 30 dB cap on the receive SINR (SINR greater than 30 dB is regarded as 30 dB). Other parameters used in the simulation are given in Table I, and are compliant with the LTE standard⁵ [32]. The results for actual packet delay with the utilization service rate model are obtained through a packet-level simulator, which adapts the transmission time of each packet to the instantaneous active interfering APs.

⁵Note that the pathloss models for different RATs need not be the same.



(a) Spectrum allocation pattern for LTE

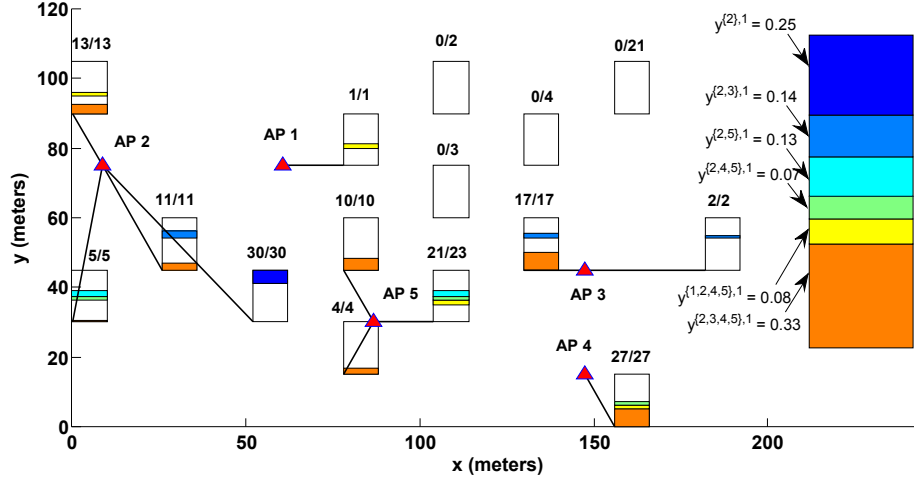


(b) Spectrum allocation pattern for LTE-U

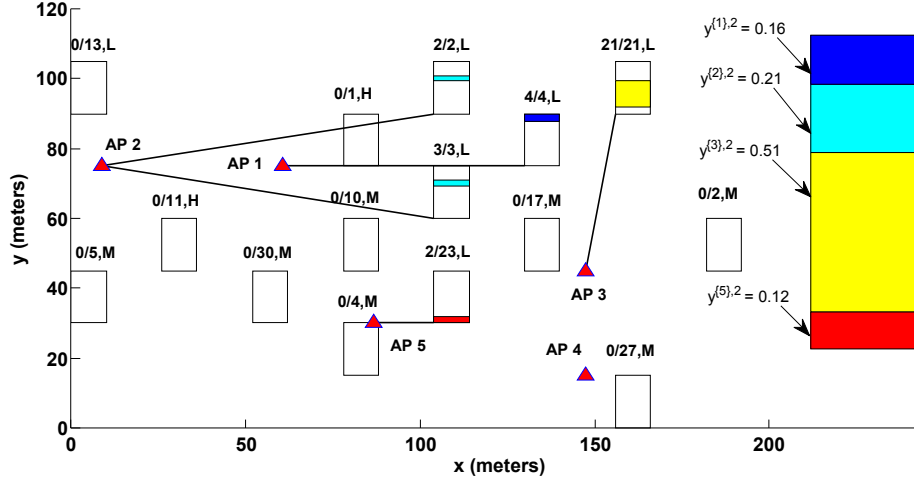
Fig. 4. Spectrum allocation patterns under heavy traffic, $\nu_H = 1$, $\nu_M = 0.01$, $\nu_L = 0.0025$.

A. The Conservative Allocation Scheme

The conservative spectrum allocation and UE-AP association according to the solution of P1 under different traffic loads are shown in Figs. 4 and 5. Each smaller rectangle indicates the spectrum allocation at the corresponding UE group. The different colors represent different patterns. The amount of spectrum resources allocated to each user group under each pattern is denoted by the size of the corresponding bar inside the rectangle. The normalized average packet arrival rate relative to the total (sum) average arrival rate of each user group as well as



(a) Spectrum allocation pattern for LTE



(b) Spectrum allocation pattern for LTE-U

Fig. 5. Spectrum allocation patterns under light/medium traffic, $\nu_H = 1, \nu_M = 0.01, \nu_L = 0.0025$.

the intensity of interference nearby from the unlicensed band are shown above the rectangle. For example, for the UE group on the top left, 19/28 of its traffic load is assigned to be served using LTE while the remaining 9/28 is to be served using LTE-U. There are three interference levels, referred to as “High”, “Medium” and “Low”, according to the intensity of the interference nearby from the unlicensed band. For concreteness, we let the corresponding second moments of the vacation duration in (8) be $\nu_H = 1, \nu_M = 0.01, \nu_L = 0.0025$. When the second moment $\nu_L = 0.0025$, the average vacation time is similar to the average packet service time, which is

about 0.05 seconds/packet. Each line segment joining an AP and a user group describes their association. The color bars on the right of each figure shows the actual spectrum partition into different patterns. Theorem 1 and Theorem 2 can be verified by counting the number of pieces in the partition and the number of rectangles connecting multiple triangles, respectively. For instance, in Fig. 4, the licensed band is partitioned into 7 segments and the unlicensed band is partitioned into 6 segments, both numbers of partitions are smaller than the number of UE groups. Moreover, there is only one UE group served by multiple APs in the unlicensed band.

Fig. 4 shows the spectrum allocation patterns for the licensed band and unlicensed band in a high traffic scenario. The licensed band is partitioned into 7 segments and the unlicensed band is partitioned into 6 segments. As we can see, the spectrum resources allocated to each user group is roughly proportional to the corresponding traffic demand. Since the licensed band alone cannot support all the UE groups, some UE groups with low traffic demand are served by using the unlicensed band only and most licensed spectrum is allocated to the UE groups with high traffic. In addition, even though some UE groups have high traffic demand (e.g., the top right one), more unlicensed spectrum is allocated to them because they do not have much nearby interference from other RATs in the unlicensed band. In contrast, some UE groups with low traffic demand (e.g., the one in the middle with arrival rate 2 packets/second) may also be allocated licensed spectrum if the nearby interference from the unlicensed band is severe.

Fig. 5 shows the spectrum allocation patterns for the licensed and unlicensed bands in a medium/low traffic scenario. The licensed band is partitioned into 6 segments and the unlicensed band is partitioned into 4 segments. Since the licensed band alone is almost enough to support all the UE groups, most of them are served by the licensed band, and only a few of the UE groups are served by the unlicensed band because of the low traffic demand and lower interference generated within that band. In addition, most spectrum is allocated to the UE groups with high traffic.

B. The Utilization-Dependent Allocation Scheme

Clearly, the delay obtained by using the conservative allocation is an upper bound on the actual delay obtained based on the utilization-dependent model. To validate the utilization-dependent model, a packet-level simulator is used to compare the theoretical delay with the actual delay. Fig. 6 compares theoretical and actual delays and service times using the utilization-dependent

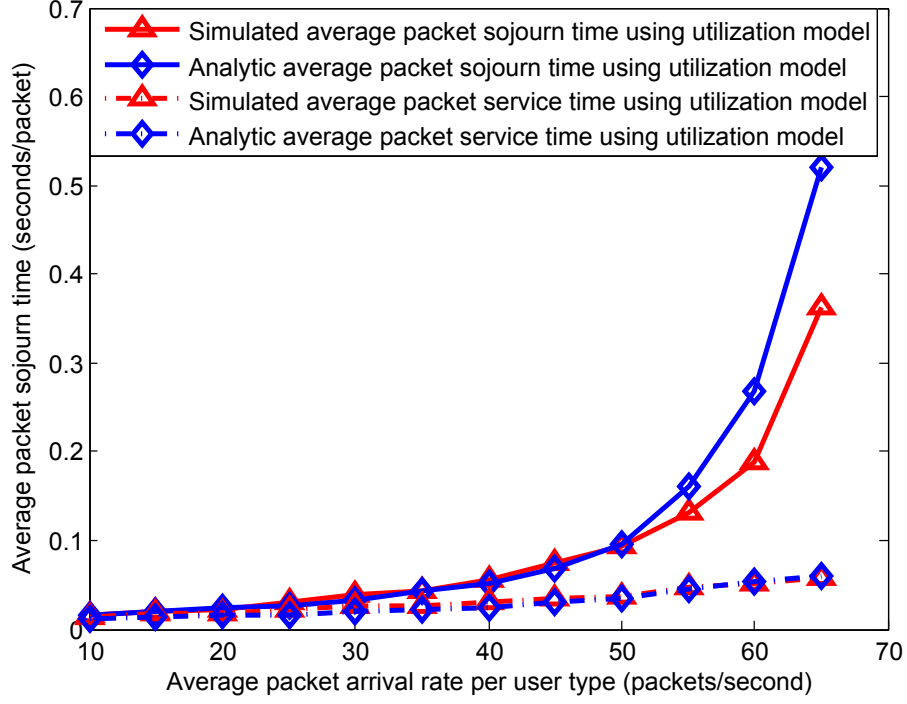


Fig. 6. Analytic and simulated delay for the utilization model.

model. The service transmission time of a packet is part of the packet sojourn time excluding the queueing delay. The horizontal axis is the average traffic load of each UE group. All curves in Fig. 6 are based on the same spectrum allocation. The utilization approximation is quite accurate with light traffic. When traffic loads are heavy, the approximation becomes coarser due to more interactions among APs. In addition, compared with delay, the service time increases much more slowly with traffic load, indicating that the spectrum is efficiently allocated to mitigate interference among APs.

Fig. 7 compares the optimization objective, average packet sojourn time, for the conservative and the utilization-dependent allocation schemes with orthogonal and full-reuse allocations. As the traffic increases to 40 and 50 packets/s, respectively, full-reuse and orthogonal allocations fail to support all the UE groups, suggesting these are the maximum throughputs under these two allocation schemes, respectively. Hence, the proposed conservative and utilization-dependent allocation schemes achieve a significantly larger throughput region than the other schemes. In addition, both conservative and the utilization-dependent allocation schemes yield significant

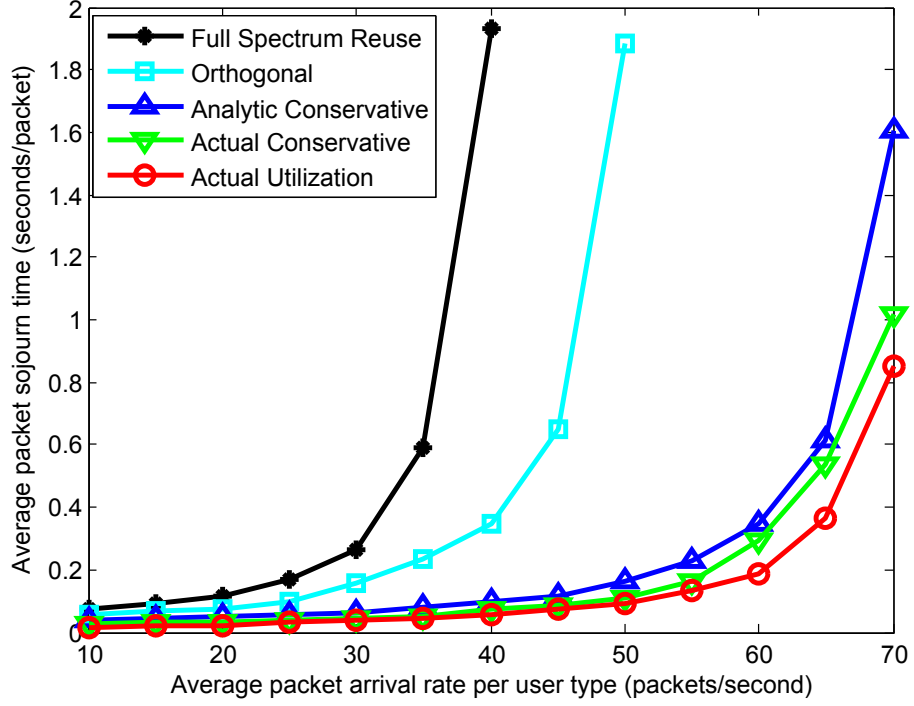
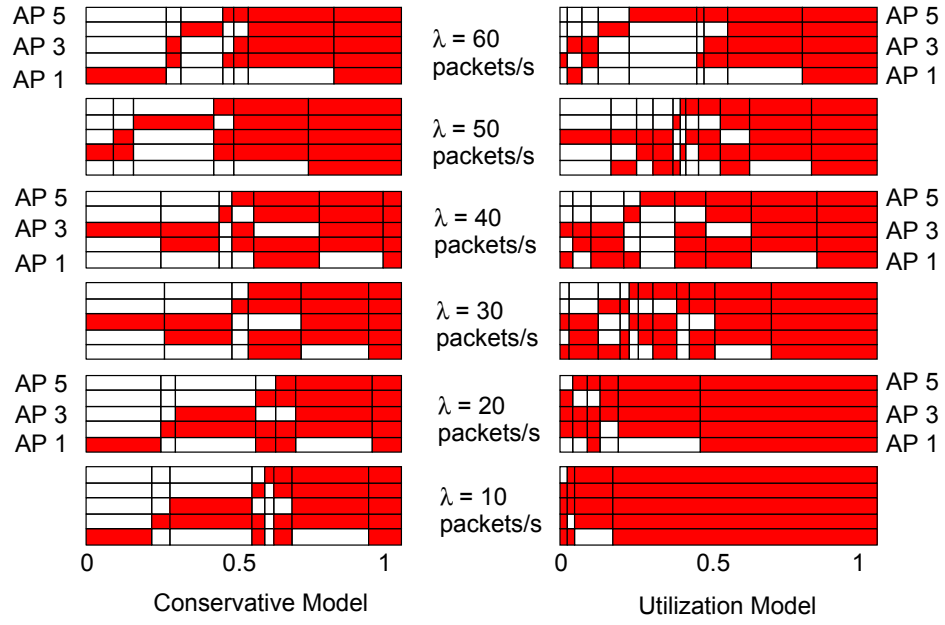


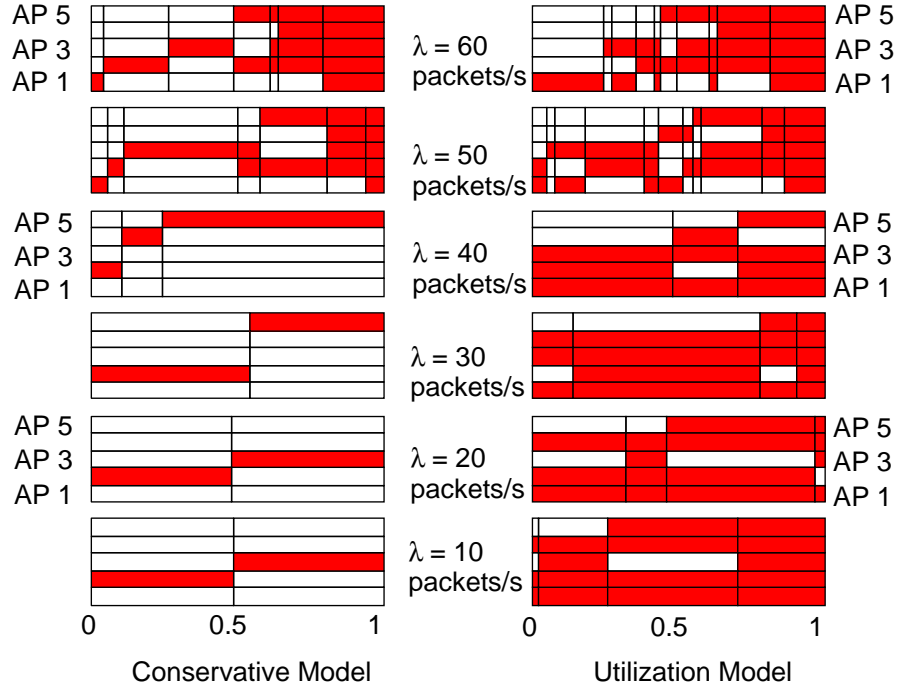
Fig. 7. Comparison of the proposed schemes with benchmark schemes.

gain especially in the high traffic scenario. The reason is that the proposed spectrum allocation schemes are traffic-aware, i.e., in the low traffic regime, it is better to reuse most of the spectrum, while in the high traffic regime, it is better to adaptively allocate the spectrum across APs. Furthermore, the proposed schemes exploit the particular characteristics of the RATs. That is, UE groups receiving less external interference from the unlicensed band are more likely to be allocated spectrum from that band. Furthermore, the actual delay based on the conservative allocation is also given, which is between the theoretical delay using conservative allocation and the actual delay using the utilization-dependent allocation scheme. The utilization-dependent allocation scheme always outperforms the conservative allocation scheme as it accurately accounts for dynamic inter-cell interference. In low traffic, the allocation based on the utilization-dependent model reduces the average delay by about 40% compared to the conservative allocation, and this improvement becomes smaller as traffic loads increase.

Fig. 8 shows the optimal allocations for the licensed and unlicensed bands under different traffic loads using the conservative and utilization-dependent allocation schemes. Rectangles



(a) Spectrum allocation patterns for LTE.



(b) Spectrum allocation patterns for LTE-U.

Fig. 8. Spectrum allocation patterns for different loads.

represent frequency bands and solid ones are used by the corresponding APs. A stack of solid rectangles then represents a reuse pattern. The length of each rectangle is proportional to the fraction of the whole spectrum occupied by the corresponding pattern. In both Fig. 8a and Fig. 8b, when traffic loads are low, the allocation scheme using the utilization-dependent model tends to be full reuse, while as traffic loads increase, it tends to orthogonalize the spectrum. In comparison with the conservative allocation, which remains orthogonal under different traffic loads, the utilization-dependent allocation uses the spectrum more efficiently since it approximates interaction among APs more accurately.

VII. CONCLUSION

We have studied spectrum allocation in downlink HetNets with multiple RATs over different bands using the average packet sojourn time as the performance metric. In addition to the licensed band, a queueing model with vacation has been proposed to model the additional delay of the unlicensed band. Two optimization based schemes have been proposed and shown to be highly effective. A sparse allocation (in terms of both the number of spectrum segments and user association) achieves the optimal network utility. Simulation results show that the proposed allocation schemes yield user associations and spectrum allocations, which utilize the spectrum of each RAT more efficiently compared with the benchmark allocation schemes, namely, orthogonal frequency reuse and full-frequency-reuse. Ongoing work is to extend the optimization framework to give a resource allocation solution with complexity that scales with the size of the network.

APPENDIX A

PROOF OF PROPOSITION 1

To analyze the queueing model with vacation and nonexhaustive service, we follow a similar technique as in [31]. Denote the residual service time at time t by $R(t)$, and the waiting and service times for the i -th packet as W_i and X_i , respectively. We have,

$$E[X] = \frac{1}{r}. \quad (17)$$

Assuming the queue is stable, by the Pollaczek-Khinchine formula,

$$E[W] = \frac{E[R]}{1 - \rho}, \quad (18)$$

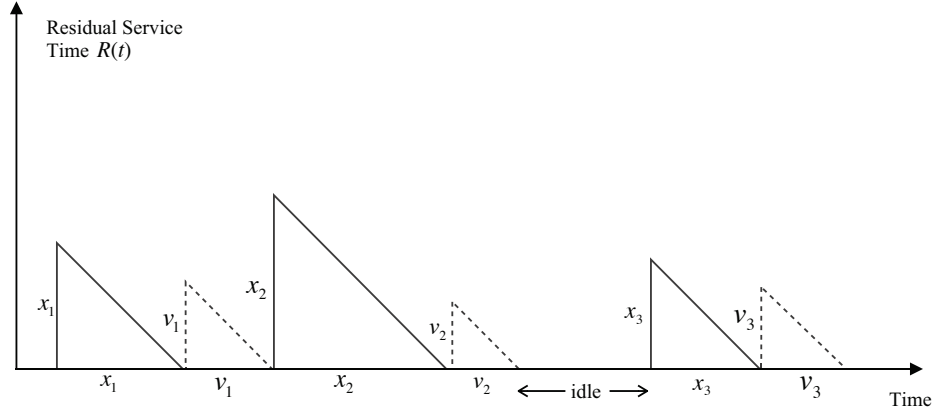


Fig. 9. Residual service time $R(t)$.

and the average packet delay is given by:

$$E[T] = E[W] + E[X] = \frac{E[R]}{1 - \rho} + \frac{1}{r}, \quad (19)$$

where $\rho = \frac{\lambda}{r}$. Denote $M(t)$ as the number of packets served up until time t . To calculate $E[R]$, we can see from Fig. 9,

$$E[R] = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t R(\tau) d\tau \quad (20)$$

$$= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{M(t)} \left(\frac{(X_i)^2}{2} + \frac{(V_i)^2}{2} \right) \quad (21)$$

$$= \lim_{t \rightarrow \infty} \frac{M(t)}{2t} \sum_{i=1}^{M(t)} \frac{(X_i)^2 + (V_i)^2}{M(t)} \quad (22)$$

$$= \frac{\lambda}{2} (E[X^2] + E[V^2]) \quad (23)$$

$$= \frac{\lambda}{r^2} + \frac{\lambda\nu}{2}. \quad (24)$$

The equalities hold in the almost sure sense. Substituting (24) into (19), we obtain (7).

To prove the convexity of (7), note that (7) can be written as the addition of two parts,

$$t = \frac{1}{(r - \lambda)^+} + \frac{\nu}{2} \frac{r\lambda}{(r - \lambda)^+}. \quad (25)$$

Since both parts are convex in the pair (r, λ) on \mathbb{R}^2 , the sum is also convex in the pair (r, λ) on \mathbb{R}^2 . ■

APPENDIX B

PROOF OF THEOREM 1

We first reformulate P1 by a change of variables. The proof then follows a similar geometric argument as in [17].

Consider a reformulation of P1 by replacing constraints (P1c), (P1e), and (P1g) with the following three constraints, introducing a new collection of variables $(z_{i \rightarrow j}^{\mathcal{A},l})$:

$$r_j^{(l)} = \sum_{\mathcal{A} \subset \mathcal{N}} \left(\sum_{i \in \mathcal{N}} s_{i \rightarrow j}^{\mathcal{A},l} z_{i \rightarrow j}^{\mathcal{A},l} \right) y^{\mathcal{A},l}, \quad j \in \mathcal{K}, l \in \{1, 2\} \quad (\text{P4a})$$

$$\sum_{j \in \mathcal{K}} z_{i \rightarrow j}^{\mathcal{A},l} \leq 1, \quad l \in \{1, 2\}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N} \quad (\text{P4b})$$

$$z_{i \rightarrow j}^{\mathcal{A},l}, y^{\mathcal{A},l} \geq 0, \quad j \in \mathcal{K}, l \in \{1, 2\}, i \in \mathcal{N}, \mathcal{A} \subset \mathcal{N}. \quad (\text{P4c})$$

The new problem, referred to as P4, is equivalent to P1. This is because the feasible set for the rate tuple \mathbf{r} remains the same, which is easy to see by regarding $z_{i \rightarrow j}^{\mathcal{A},l}$ as the fraction of spectrum under pattern \mathcal{A} that AP i allocates to UE group j over RAT l . If one solves P4, the actual spectrum allocations can be recovered as $x_{i \rightarrow j}^{\mathcal{A},l} = y^{\mathcal{A},l} z_{i \rightarrow j}^{\mathcal{A},l}$.

We show that if \mathbf{r}^* is an optimal rate tuple of P1, then we can attain each sub-tuple $\mathbf{r}^{*(l)}$ over RAT l with a k -sparse $\mathbf{y}^{(l)}$. For each RAT l , let us begin with a generally nonsparse optimal solution $\mathbf{y}^{(l)}$, whose support is $\mathcal{S}^{(l)}$ ($y^{\mathcal{A},l} = 0$ if $\mathcal{A} \notin \mathcal{S}^{(l)}$). Let us freeze the optimal $z_{i \rightarrow j}^{\mathcal{A},l}$ variables and define a k -vector $\mathbf{q}^{\mathcal{A},l}$ for every $\mathcal{A} \in \mathcal{S}^{(l)}$ with its elements determined by $q_j^{\mathcal{A},l} = \sum_{i \in \mathcal{N}} s_{i \rightarrow j}^{\mathcal{A},l} z_{i \rightarrow j}^{\mathcal{A},l}$. According to (P4a), a convex combination of the vectors $(\mathbf{q}^{\mathcal{A},l})_{\mathcal{A} \in \mathcal{S}^{(l)}}$ with $(y^{\mathcal{A},l})_{\mathcal{A} \in \mathcal{S}^{(l)}}$ as coefficients form the optimal rate tuple $\mathbf{r}^{(l)}$ over RAT l . By Carathéodory's Theorem [33], $\mathbf{r}^{(l)}$ can be represented as a convex combination of at most $k + 1$ of those vectors. Moreover, $\mathbf{r}^{(l)}$ must be on the boundary, not in the interior of the convex hull of $(\mathbf{q}^{\mathcal{A},l})$, because otherwise the rate tuple can be increased in all dimensions, contradicting the optimality assumption. Thus, for each RAT l , we can identify a vector $\mathbf{y}^{*(l)}$ whose support is a subset of $\mathcal{S}^{(l)}$ with k or fewer elements, such that the optimal $\mathbf{r}^{(l)}$ is a convex combination of $(\mathbf{q}^{\mathcal{A},l})$ with $y^{*\mathcal{A},l}$ as coefficients. ■

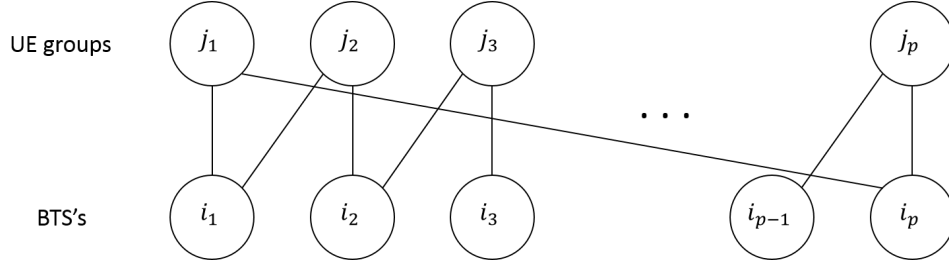


Fig. 10. A loop in the BGR for the UE groups served by multiple APs over the same RAT.

APPENDIX C

PROOF OF THEOREM 2

The proof follows an analogous proof in [5]. The KKT conditions for P1 for nonzero elements of \mathbf{x} and \mathbf{y} are :

$$\frac{\partial U}{\partial r_j^{(l)}} s_{i \rightarrow j}^{\mathcal{A},l} - \mu_i^{\mathcal{A},l} = 0, \quad (26a)$$

$$\sum_{i \in \mathcal{N}} \mu_i^{\mathcal{A},l} = \xi^{(l)}, \quad (26b)$$

where $\mu_i^{\mathcal{A},l}$ and $\xi^{(l)}$ are the Lagrange multipliers for constraints (P1e) and (P1f), respectively. Assuming UE group j is served by two APs i_1 and i_2 over RAT l , define $s_{i \rightarrow j}^{(l)} = \sum_{\mathcal{A} \subset \mathcal{N}} s_{i \rightarrow j}^{\mathcal{A},l}$ and $\mu_i^{(l)} = \sum_{\mathcal{A} \subset \mathcal{N}} \mu_i^{\mathcal{A},l}$. According to (26), we have

$$\frac{s_{i_1 \rightarrow j}^{(l)}}{s_{i_2 \rightarrow j}^{(l)}} = \frac{\mu_{i_1}^{(l)}}{\mu_{i_2}^{(l)}}. \quad (27)$$

Following the argument in [5], a bipartite graph representation is used as in [34]. For each RAT, denote the UE groups served by multiple APs and the corresponding APs as nodes. An edge between a UE group and an AP represents an association. It remains to show that the graph contains no loop. Suppose there is a loop in the bipartite graph as shown in Fig. 10. Then there exists a sequence of nodes of users j_1, \dots, j_p and a sequence of APs i_1, \dots, i_p , where user j_q is connected with AP i_q for $q = 1, \dots, p$, user j_{q+1} is connected with AP i_q for $q = 1, \dots, p-1$, and user 1 is connected with AP p . The nodes are distinct otherwise we can find a smaller loop with this property. According to (27), the loop implies:

$$\frac{s_{i_p \rightarrow j_1}^{(l)}}{s_{i_1 \rightarrow j_1}^{(l)}} \frac{s_{i_1 \rightarrow j_2}^{(l)}}{s_{i_2 \rightarrow j_2}^{(l)}} \dots \frac{s_{i_{p-1} \rightarrow j_p}^{(l)}}{s_{i_p \rightarrow j_p}^{(l)}} = \frac{\mu_{i_p}^{(l)}}{\mu_{i_1}^{(l)}} \frac{\mu_{i_1}^{(l)}}{\mu_{i_2}^{(l)}} \dots \frac{\mu_{i_{p-1}}^{(l)}}{\mu_{i_p}^{(l)}} = 1.$$

Since $s_{i \rightarrow j}^{(l)}$ is the sum of spectral efficiencies over all patterns over RAT l , it is a random variable based on random topology. Therefore, $\frac{s_{ip \rightarrow j1}^{(l)} s_{i1 \rightarrow j2}^{(l)}}{s_{i1 \rightarrow j1}^{(l)} s_{i2 \rightarrow j2}^{(l)}} \dots \frac{s_{ip-1 \rightarrow jp}^{(l)}}{s_{ip \rightarrow jp}^{(l)}} = 1$ is a zero probability event, which shows that w.p.1 there is no loop in the graph. Since there are n APs, the largest possible BGR without a loop has $n - 1$ user nodes, which proves Theorem 2. ■

APPENDIX D

PROOF OF PROPOSITION 2

Taking the second derivatives of $\lambda_j^{(1)} \hat{t}_j^{(1)}$ with respect to $\lambda_j^{(1)}$ and $r_j^{(1)}$, we have:

$$\frac{\partial^2 \lambda_j^{(1)} \hat{t}_j^{(1)}}{\partial \lambda_j^{(1)2}} = \frac{\eta r_j^{(1)}}{(r_j^{(1)} - \beta \lambda_j^{(1)})^3}, \quad (28a)$$

$$\frac{\partial^2 \lambda_j^{(1)} \hat{t}_j^{(1)}}{\partial r_j^{(1)2}} = \frac{\eta \lambda_j^{(1)}}{\beta} \left(\frac{1}{(r_j^{(1)} - \beta \lambda_j^{(1)})^3} - \frac{1}{(r_j^{(1)})^3} \right) + \frac{2\beta \lambda_j^{(1)}}{(r_j^{(1)})^3}. \quad (28b)$$

When $\frac{1}{\lambda_j^{(1)}} > \frac{\beta}{r_j^{(1)}}$, in other words, when the queue is stable, the derivatives are positive, which means $\lambda_j^{(1)} \hat{t}_j^{(1)}$ is bi-convex in $\lambda_j^{(1)}$ and $r_j^{(1)}$. In addition, since $\lambda_j^{(2)} \hat{t}_j^{(2)}$ is similar to $\lambda_j^{(1)} \hat{t}_j^{(1)}$ except for an additional term $\frac{\nu_j(\lambda_j^{(2)})^2 r_j^{(2)}}{2(r_j^{(2)} - \beta \lambda_j^{(2)})^+}$, which is also bi-convex in $\lambda_j^{(2)}$ and $r_j^{(2)}$, $\lambda_j^{(2)} \hat{t}_j^{(2)}$ is also bi-convex in $\lambda_j^{(2)}$ and $r_j^{(2)}$. Since \hat{U} is a linear combination of bi-convex functions, we conclude that (17) is bi-convex in λ and r . ■

REFERENCES

- [1] Z. Zhou, D. Guo and M. L. Honig, "Allocation of licensed and unlicensed spectrum in heterogeneous networks", in *Proc. IEEE Australian Communication Theory Workshop*, Melbourne, VIC, Australia, 2016.
- [2] A. Khandekar, N. Bhushan, T. Ji, and V. Vanghi, "LTE-Advanced: Heterogeneous networks," in *European Wireless Conference*, pp. 978-982, 2010.
- [3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065-1082, 2014.
- [4] Q. Kuang and W. Utschick, "Energy management in heterogeneous networks with cell activation, user association and interference coordination," *IEEE Trans. Wireless Commun.*, vol. 15, pp. 3868-3879, 2016.
- [5] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and multi-pattern resource allocation in heterogeneous networks," *IEEE Trans. Sig. Proc.*, vol. 64, pp. 3388-3401, 2016.
- [6] A. L. Stolyar and H. Viswanathan, "Self-organizing dynamic fractional frequency reuse in OFDMA systems," *IEEE INFOCOM*, Phoenix, USA, Apr. 2008.
- [7] R. Chang, Z. Tao, J. Zhang, and C.-C. Kuo, "Multicell OFDMA downlink resource allocation using a graphic framework," *IEEE Trans. Veh. Technol.*, vol. 58, pp. 3494-3507, Sep. 2009.

- [8] S. H. Ali and V. C. M. Leung, "Dynamic frequency allocation in fractional frequency reused OFDMA networks," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 4286-4295, Aug. 2009.
- [9] R. Madan, J. Borran, A. Sampath, N. Bhushan, A. Khandekar, and T. Ji, "Cell association and interference coordination in heterogeneous LTE-A cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1479-1489, 2010.
- [10] W.-C. Liao, M. Hong, Y.-F. Liu, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939-3952, 2014.
- [11] R. Etkin, A. Parekh and D. Tse, "Spectrum sharing for unlicensed bands," in *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 517-528, April 2007.
- [12] J. Huang, V. G. Subramanian, R. A. Agrawal, and R. Berry, "Joint scheduling and resource allocation in uplink OFDM systems for broadband wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 27, pp. 226-234, Feb. 2009.
- [13] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248-257, 2013.
- [14] G. Lim, C. Xiong, L. J. Cimini Jr., and G. Y. Li, "Energy-efficient resource allocation for OFDMA-based multi-RAT networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2696-2705, 2014.
- [15] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1100-1113, 2014.
- [16] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets," *IEEE/ACM Trans. Netw.*, vol. 22, no. 1, pp. 137-150, 2014.
- [17] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun. Special Issue on Recent Advances in Heterogeneous Cellular Networks*, vol. 33, pp. 2027-2038, 2015.
- [18] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun. Special Issue on Energy-Efficient Techniques for 5G Wireless Communication Systems*, vol. 34, pp. 823-831, 2016.
- [19] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, "Cell-free massive MIMO versus small cells," <http://arxiv.org/abs/1602.08232>, 2016.
- [20] F. Teng and D. Guo, "Resource management in 5G: A tale of two timescales," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2015.
- [21] F. Teng and D. Guo, "Dual-timescale spectrum management in wireless heterogeneous networks," <http://arxiv.org/abs/1604.02781>, 2016.
- [22] F. Liu, E. Bala, E. Erkip, M. C. Beluri and R. Yang, "Small-cell traffic balancing over licensed and unlicensed bands," in *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5850-5865, Dec. 2015.
- [23] A. R. Elsherif, W. P. Chen, A. Ito, and Z. Ding, "Resource allocation and inter-cell interference management for dual-access small cells," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1082-1096, Jun. 2015.
- [24] Q. Chen, G. Yu, R. Yin, A. Maaref, G. Y. Li, and A. Huang, "Energy efficiency optimization in licensed-assisted access," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, April 2016.
- [25] C. Godsil and G. Royle, *Algebraic Graph Theory*. New York: Springer-Verlag, 2001.
- [26] N. Jindal and D. Breslin, "LTE and Wi-Fi in unlicensed spectrum: a coexistence study," *White paper*, 2015.
- [27] R. Ratasuk, M. A. Uusitalo, N. Mangalvedhe, A. Sorri, S. Iraj, C. Wijting, and A. Ghosh, "License-exempt LTE deployment in heterogeneous network," in *Proc. 9th International Symposium on Wireless Communication Systems*, Paris, France, Aug. 2012.

- [28] F. S. Chaves, E. P. L. Almeida, R. D. Vieira, A. M. Cavalcante, F. M. Abinader Jr., S. Choudhury, and K. Doppler. "LTE UL power control for the improvement of LTE/Wi-Fi coexistence." *IEEE Vehicular Technology Conference*, pp. 1-6, 2013.
- [29] E. P. L. Almeida, A. M. Cavalcante, R. C. D. Paiva, F. S. Chaves, F. M. Abinader Jr., R. D. Vieira, S. Choudhury, E. Tuomaala, and K. Doppler, "Enabling LTE/WiFi coexistence by LTE blank subframe allocation," *IEEE International Communications Conference*, pp. 1-6, 2013.
- [30] T. Nihtila, V. Tykhomyrov, O. Alanen, M. Uusitalo, A. Sorri, M. Moisio, S. Irja, R. Ratasuk, and N. Mangalvedhe, "System performance of LTE and IEEE 802.11 coexisting on a shared frequency band," in *Proc. IEEE Wireless Communications and Networking Conf.*, Shanghai, China, pp. 1056-1061, Apr. 2013.
- [31] N. Tian and Z. G. Zhang, *Vacation queueing models: theory and applications*. Springer, 2006.
- [32] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," V9.0.0, March 2010.
- [33] H. G. Eggleston, *Convexity*. No. 47, Cambridge University Press Archive, 1958.
- [34] V. Gajic, J. Huang, and B. Rimoldi, "Competition of wireless providers for atomic users: Equilibrium and social optimality," in *Proc. Allerton Conf. Commun., Control, & Computing*, pp. 1203-1210, Monticello, IL, USA, 2009.